When Variety-Seeking Meets Unexpectedness: Incorporating Variety-Seeking Behaviors into Design of Unexpected Recommender Systems

Appendix

Part I: The Curiosity Questionnaire and Statistics of Consumer Response

Questions	Response Scale
Q1: "I actively seek as much information as I can in new situations."	7-point Likert scale 1–strongly disagree 2–moderately disagree 3-disagree a little 4–neither agree nor disagree 5-agree a little 6–moderately agree 7–strongly agree
Q2: "I am the type of person who really enjoys the uncertainty of everyday life."	
Q3: "I am at my best when doing something that is complex or challenging."	
Q4: "Everywhere I go, I am out looking for new things or experiences."	
Q5: "I view challenging situations as an opportunity to grow and learn."	
Q6: "I like to do things that are a little frightening."	
Q7: "I am always looking for experiences that challenge how I think about myself and the world."	
Q8: "I prefer jobs that are excitingly unpredictable."	
Q9: "I frequently seek out opportunities to challenge myself and grow as a person."	
Q10: "I am the kind of person who embraces unfamiliar people, events, and places."	

Table 1: "Ten-item Curiosity and Exploration Inventory-II" Questionnaire (Kashdan et al. 2009)

Survey Question & Response	Mean	Std.	Median	Skewness	Kurtosis
"The item recommended to me matches my interests."	3.32	1.410	4.00	-0.419	-1.192
The item recommended to me is novel."	3.06	1.424	3.00	-0.146	-1.391
"The item recommended to me is different from the types of products I bought before."	3.39	1.215	4.00	-0.400	-0.813
The item recommended to me is similar to the system's prior recommendations."	2.93	1.302	3.00	0.214	-1.109
"The item recommended to me is unexpected."	3.16	1.437	3.00	-0.199	-1.337
"The item recommended to me is a pleasant surprise."	2.73	1.456	2.50	0.195	-1.400
"The item recommended to me is very timely."	3.00	1.484	3.00	-0.074	-1.450
"I am satisfied with this recommendation."	3.21	1.140	3.00	-0.286	-0.466
"I would buy the item recommended, given the opportunity."	2.83	1.456	3.00	0.003	-1.418
"Ten-item Curiosity and Exploration Inventory-II"	3.13	0.831	3.10	0.088	-0.402

Table 2: Statistics of the Consumer Responses to the Questionnaire

Part II: Classification Performance on the Curiosity Questionnaire

Variety-Seeking Framework	Accuracy	F1-Score
Euclidean+Exponential+Mean **(%Improved)**	**0.937***** **(0.017)** **+%**	**0.867***** **(0.012)** **+%**
Euclidean+Hyperbolic+Mean	0.919***	0.839***
Euclidean+No Decay+Mean	0.790***	0.749***
Cosine+Exponential+Mean	0.908***	0.837***
Cosine+Hyperbolic+Mean	0.901***	0.808***
Cosine+No Decay+Mean	0.779**	0.716***
Manhattan+Exponential+Mean	0.873***	0.825***
Manhattan+Hyperbolic+Mean	0.864***	0.801***
Manhattan+No Decay+Mean	0.776**	0.709***
Chebyshev+Exponential+Mean	0.877***	0.784***
Chebyshev+Hyperbolic+Mean	0.865***	0.763***
Chebyshev+No Decay+Mean	0.765	0.682***
Feature+Exponential+Mean	0.832***	0.726***
Feature+Hyperbolic+Mean	0.837***	0.721***
Feature+No Decay+Mean	0.761	0.643
Binary+Exponential+Mean	0.778**	0.668**
Binary+Hyperbolic+Mean	0.771**	0.641
Binary+No Decay+Mean	0.738	0.640
APF	0.730	0.629
PAS	0.763	0.639

*Table 3: Classification Performance of Consistency- v.s. Variety-Seeking Consumers Based on Our Variety-Seeking Framework. The threshold is the average variety-seeking level across all consumers. *p<0.1; **p<0.05; ***p<0.01. (compared to APF & PAS)*

Part III: Analysis of the Statistics to Use for the Stationarity Assumption

Variety-Seeking Framework	$Product\_Variety(i,j,t)$	$Variety\_Seeking(i)$
Euclidean+Exponential+Arithmetic Mean	**0.775**** **(0.004)**	**0.618**** **(0.003)**
Euclidean+Exponential+Weighted Mean	0.773	0.616
Euclidean+Exponential+Geometric Mean	0.771	0.613
Euclidean+Exponential+Harmonic Mean	0.770	0.613
Euclidean+Exponential+Median	0.728	0.589

*Table 4: Pearson Correlation Coefficients between Self-Reported Variety-Seeking Levels and Our Variety-Seeking Framework. *p<0.1; **p<0.05; ***p<0.01. (compared to APF & PAS)*

Part IV: Summary Statistics of Consumer and Video Features

Consumer /Video Features		Mean	25th percentile	Median	75th percentile	Variance
Gender	Treatment	0.50025	0.000	1.000	1.000	0.495
	Control	0.50015	0.000	1.000	1.000	0.495
Age	Treatment	41.119	30.000	40.000	50.000	21.100

	Control	40.076	30.000	40.000	50.000	20.070
VIP Status	Treatment	0.361	0.000	0.000	1.000	0.480
	Control	0.365	0.000	0.000	1.000	0.480
Activity Days	Treatment	20.668	8.000	19.000	27.000	8.520
	Control	20.404	8.000	19.000	27.000	8.450
Genre		4.731	1.000	4.000	6.000	2.150
View Count		9.684	8.000	10.000	12.000	3.128
Comment Count		7.070	0.000	2.000	5.000	4.283
Release Days		6.325	5.000	7.000	8.000	2.304

Table 5: Summary statistics of consumer and video features for the control and treatment groups.

Part V: Additional Results in the Online Experiment: Difference-in-Difference Analysis and User-Level Analysis

To further justify the validity of our empirical findings, we replicate our analysis in Section 5.3 using the Difference-in-Difference method instead, where we specify the ATEs in equation (4) of the paper using the following alternative identification:

$$Metric_{ij} = \alpha_0 + \alpha_1 * Treatment_i + \alpha_2 * Experiment_t + \alpha_3 * Treatment_i * Experiment_t + \overrightarrow{\alpha_4} * \overrightarrow{X_i} + \overrightarrow{\alpha_5} * \overrightarrow{Y_j} + \varepsilon_{ij}$$

where $Metric_{ij} \in \{CTR_{ij}, VV_{ij}, TS_{ij}\}$, $Treatment_i$ is the dummy variable indicating whether consumer i is in the treatment group or not, $Experiment_t$ is the dummy variable indicating the post-treatment period versus the pre-treatment period, $\overrightarrow{X_i}$ represents explicit user features, and $\overrightarrow{Y_j}$ represents explicit video features. We can observe from Table 6 almost identical treatment effects as we report in Section 5.3, as our proposed model achieves significant performance improvements on those consumers in the treatment group. Besides, we observe that $Experiment_t$ does not hold a significant coefficient, suggesting that the parallel trends assumption is fulfilled and that the observed relationship is unlikely to arise as an artifact from events occurred before our treatment.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229*** (0.0042)	0.0455*** (0.0028)	39.210*** (0.9891)
User Features Video Features	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0079	0.0043	0.2126
Observations	46,794,473	46,794,473	46,794,473

*Table 6: Average Treatment Effect Using the Difference-in-Difference Analysis. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Furthermore, in addition to the three business metrics that we study in the paper, we also analyze the treatment effects on the following user-level metrics in this section: (a) CTR_i, the average click-through rate of user i in each session; (b) VV_i, the average finish-watching percentage of user i in each session, and (c) TS (Time Spent), the average time the user has spent in each session. We specify the ATEs using the following identification:

$$Metric_i = \alpha_0 + \alpha_1 * Treatment_i + \overrightarrow{\alpha_2} * \overrightarrow{X_i} + D_t + \varepsilon_{ij}$$

where $Metric_{ij} \in \{CTR_i, VV_i, TS_i\}$, $Treatment_i$ is the dummy variable, $\overrightarrow{X_i}$ represents explicit user features, and D_t represents time-fixed effects including dates and hours. We can observe from Table 7 that consumers who are served by our proposed model achieve significant improvements in all three user-level metrics compared to the control group.

	CTR_i	VV_i	TS_i
$Treatment_i$	0.0227*** (0.0049)	0.0459*** (0.0033)	225.378*** (5.7478)
User Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0075	0.0041	0.2025
Observations	37,965,781	37,965,781	37,965,781

*Table 7: Average Treatment Effect on the User-Level Metrics. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Part VI: Statistics and Experiment Results of the Sparsity Analysis

Dataset	Alibaba-1	Alibaba-2	Alibaba-3
#Consumers	46,143	21,152	10,737
#Products	53,657	22,809	13,751
#Transactions	1,806,157	1,061,948	646,933
#Records Per Consumer	39.14	50.21	60.25
Sparsity	0.073%	0.220%	0.438%

Table 8: Descriptive Statistics of the Datasets for Sparsity Analysis

	Alibaba-1		Alibaba-2		Alibaba-3	
	AUC	HR@10	AUC	HR@10	AUC	HR@10
DIN+Latent+Multiply **(%Improved)**	**0.7349***** **(0.0088)** **+2.73%**	**0.7730***** **(0.0089)** **+3.15%**	**0.7511***** **(0.0092)** **+4.62%**	**0.7802***** **(0.0097)** **+3.99%**	**0.7670***** **(0.0094)** **+6.22%**	**0.7885***** **(0.0101)** **+4.85%**
DIN+Latent+Exponential	0.7328***	0.7701***	0.7470***	0.7763***	0.7618***	0.7852***

DIN+Latent+Power	0.7324***	0.7688***	0.7461***	0.7751***	0.7599***	0.7838***
DIN+Feature+Multiply	0.7299***	0.7672***	0.7458***	0.7749***	0.7580***	0.7832***
DIN+Feature+Exponential	0.7303***	0.7654***	0.7427***	0.7738***	0.7564***	0.7815***
DIN+Feature+Power	0.7291***	0.7658***	0.7411***	0.7726***	0.7573***	0.7806***
NCF+Latent+Multiply	0.7266**	0.7649***	0.7409***	0.7730***	0.7562***	0.7802***
NCF+Latent+Exponential	0.7261**	0.7610**	0.7388***	0.7722***	0.7558***	0.7794***
NCF+Latent+Power	0.7249**	0.7587**	0.7395***	0.7719***	0.7549***	0.7799***
NCF+Feature+Multiply	0.7228**	0.7599**	0.7392***	0.7698***	0.7522***	0.7777***
NCF+Feature+Exponential	0.7237**	0.7576**	0.7376***	0.7684***	0.7515***	0.7760***
NCF+Feature+Power	0.7240**	0.7573**	0.7383***	0.7672***	0.7516***	0.7765***
DIN	0.6957	0.6972	0.7026	0.7028	0.7055	0.7047
DeepFM	0.5519	0.5164	0.5917	0.5579	0.6214	0.5892
PURS	0.7154	0.7494	0.7179	0.7503	0.7221	0.7520
HOM-LIN	0.5812	0.5493	0.6032	0.5871	0.6075	0.6061
Re-Ranking	0.6025	0.5776	0.6164	0.6011	0.6215	0.6163
DPP	0.6517	0.7026	0.6662	0.7075	0.6707	0.7124
LinUCB	0.6775	0.6662	0.6825	0.7101	0.7028	0.6925
COFIBA	0.6796	0.6798	0.6906	0.7129	0.7032	0.7016

*Table 9: Offline results on the three Alibaba datasets. *p<0.1; **p<0.05; ***p<0.01. Improvement percentages were reported over the second-best baseline models (underlined).*

Part VII: Additional Results for Robustness Check

(a) We include different combinations of consumer features, video features, and time-fixed effects in the regression model to evaluate the treatment effects of adopting our proposed model. As shown in Table 10, our results will not be affected by the particular model specification of features or fixed effects during the estimation process.

	CTR_{ij}	CTR_{ij}	CTR_{ij}	CTR_{ij}
$Treatment_i$	0.0229*** (0.0043)	0.0233*** (0.0047)	0.0240*** (0.0102)	0.0238*** (0.0129)
User & Video Features Time Fixed Effect	Yes Yes	Yes No	No Yes	No No
R-Squared	0.0081	0.0069	0.0022	0.0011
Observations	37,965,781	37,965,781	37,965,781	37,965,781

*Table 10: Average Treatment Effect with Different Combinations of Features and Fixed Effects. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01*

(b) We use alternative models to specify the binary outcome variables CTR_{ijt} and VV_{ijt}, including the discrete choice models of Logit and Probit. The results in Table 11 show that treatment effects are still positive and statistically significant under different specifications.

	CTR_{ij} (Linear)	CTR_{ij} (Logit)	CTR_{ij} (Probit)

$Treatment_i$	0.0229*** (0.0043)	0.571*** (0.076)	0.403*** (0.059)
User & Video Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0081	0.0084	0.0082
Observations	37,965,781	37,965,781	37,965,781

*Table 11: Average Treatment Effect of Different Specification Models. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

(c) We also exclude recommendation records of video content uploaded by Company A itself during the experiment, as some consumers might be more willing to click on these video recommendations due to their loyalty to the platform. As shown in Table 12, the estimation results are not significantly different from the original estimation.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229*** (0.0044)	0.0456*** (0.0028)	39.206*** (1.0003)
User & Video Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0079	0.0043	0.2126
Observations	37,146,255	37,146,255	37,146,255

*Table 12: Average Treatment Effect after Excluding Self-Uploaded Video Content. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

(d) Finally, we drop the records from those users in the regions where the platform of Company A was launched recently, as those users might be more willing to click on whatever recommendations they receive due to the novelty effect. The results demonstrated in Table 13 show that our treatment effects remain robust when we exclude those records.

	CTR_{ij}	VV_{ij}	TS_{ij}
$Treatment_i$	0.0229*** (0.0043)	0.0455*** (0.0028)	39.222*** (0.9897)
User & Video Features Time Fixed Effect	Yes Yes	Yes Yes	Yes Yes
R-Squared	0.0080	0.0045	0.2136
Observations	37,894,442	37,894,442	37,894,442

*Table 13: Average Treatment Effect after Excluding Recently Launched Records. The table shows a regression with robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Part VIII: Product-Level Analysis

As we have already demonstrated the significant impact of our proposed model on the consumer side, we will study its influence on the product side in this section, particularly on product demand

distribution (Tan et al. 2017; Fong 2017). To do this, we compare the Lorenz curve & Gini Coefficient of the clicked videos between the treatment group and the control group in our experiment. Note that the Gini Coefficients of both groups are the same over the pre-treatment period, as part of the randomized setting. As shown in Figure 1, the inequalities of video distributions in the treatment group (Gini Index=0.44) have significantly decreased, compared to those in the control group (Gini Index=0.59). This is the case, as our proposed model improves consumers' variety-seeking levels in general, resulting in more unexpected video content in recommendations, which are typically novel and have little exposure under classical recommender systems. Therefore, our model has the potential to alleviate the "winners-take-most" and fairness problems in recommendations (Wang et al. 2016) and contribute to a better consumer experience.

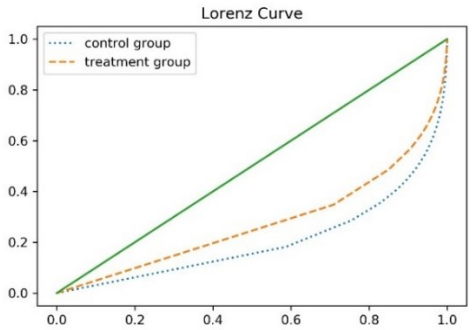


Figure 1: Lorenz Curve of the video view count in the treatment and control group

Part IX: Churn Rate Analysis

To make sure that consumer abandonment behavior (i.e., consumers might reduce their usage or leave the platform if served by our proposed model) and curiosity factors (i.e., consumers might be curious to try out the new recommender system design) do not significantly contribute to the performance improvements, we conduct additional analysis to study the long-term treatment effects and changes of the churn rate, which is measured as a binary variable on a daily basis of whether the consumer watches any video on that day or not. In particular, we specify the churn rate variable of consumer i at day t using the logistic regression:

$$Churn\_Rate_{it} = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 * Treatment_i + \overrightarrow{\alpha_2} * \overrightarrow{X_i} + D_t + \varepsilon_{it})}} \qquad (5)$$

and we also specify the long-term treatment effects through the OLS regression model:

$$Metric_{ijt} = \alpha_0 + \alpha_1 * Treatment_i * D_t + \overrightarrow{\alpha_2} * \overrightarrow{X_\iota} + \overrightarrow{\alpha_3} * \overrightarrow{Y_j} + \varepsilon_{ijt} \qquad (6)$$

where D_t represents the date in our online experiments. The regression results in Table 14 show that the churn rate will be significantly lower if consumers adopt the treatment of our proposed model, indicating that *more* users choose to stay with the platform compared to the existing model. This is the case, as our proposed model produces more satisfying video recommendations for the consumers to keep them within the platform. Therefore, we demonstrate that those significant improvements achieved by our model do not come from the abandonment effect, as our model reduces the churn rate significantly.

	$Churn\_Rate_{it}$
$Treatment_i$	-0.0611*** (0.0043)
User Features Time Fixed Effect	Yes Yes
R-Squared	0.0081
Observations	37,965,781

*Table 14: Average Treatment Effect of Variety-Seeking Based Unexpected Recommendations on Churn Rate. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Part X: Variety-Seeking Behavior Analysis

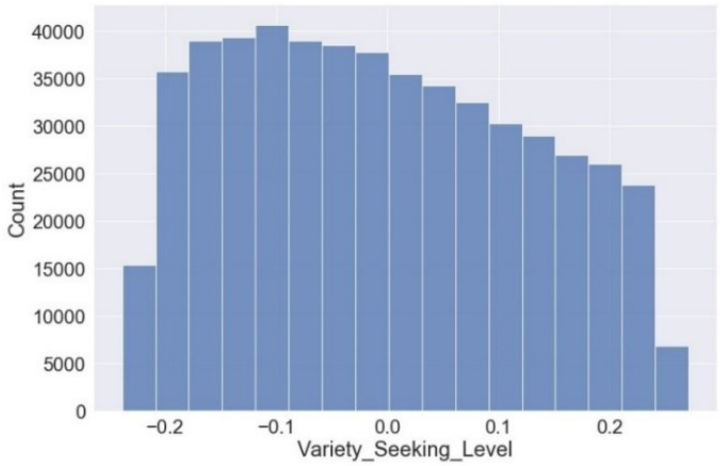
To improve our understanding of the variety-seeking behavior, we plot the pre-experiment distribution of variety-seeking levels in the treatment group in Figure 2(a), which is close to a skewed normal distribution where the majority of consumers have low variety-seeking levels. After being served by our model, their variety-seeking behaviors have significantly changed based on their experience with the new system, where we have the following observations in Figure 2(b):

First, consumers in general would seek more variety in the video recommendations after being served by our proposed method, as the average variety-seeking level $Variety\_Seekiing(i)$ in the treatment group has significantly increased from -0.0053 to 0.0232 after the adoption.

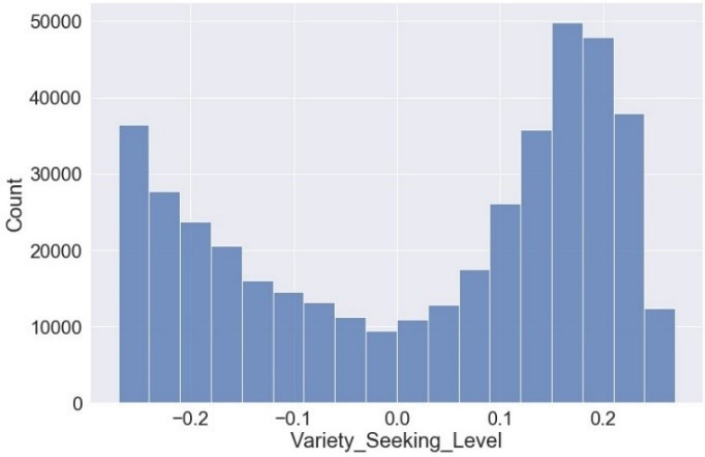
Second, our method reinforces the variety-seeking behavior for those consumers with high-level of variety-seeking behavior ($Variety(i)$>0.1), as we witness a significant increase in the average

variety-seeking level among these consumers from 0.1434 to 0.1781, an improvement of 24.20%. At the same time, our method further reduces the level of variety in the recommended videos for those consumers with low-level of variety-seeking behavior ($Variety(i)$<-0.1), a decrease of 25.53% from -0.1594 to -0.2001.

Third, our stationary assumption still holds for consumers served by our proposed model. In particular, we repeat the ADF test on post-treatment records, and the average statistics in the treatment group is -33.28, more negative than the critical value of −3.5 at the 95% confidence level. This observation further demonstrates the validity of the stationarity assumption that we make in our variety-seeking framework.



(a) Before the Adoption



(b) After the Adoption

Figure 2: Comparisons of the Variety-Seeking Levels of Consumers